Assignment 6

Ariful Islam

June 28, 2025

1:

In this experiment, we used a pre-trained MobileNetV2 model to classify an image of a **horned viper**. Initially, the model predicted it correctly with a confidence of 89.94

We then applied the Fast Gradient Sign Method (FGSM) with epsilon = 0.15 to generate an adversarial image. The new image was visually similar but misclassified as **sidewinder** with 31.39 percent confidence. This confirms that FGSM can fool a neural network using imperceptible noise.

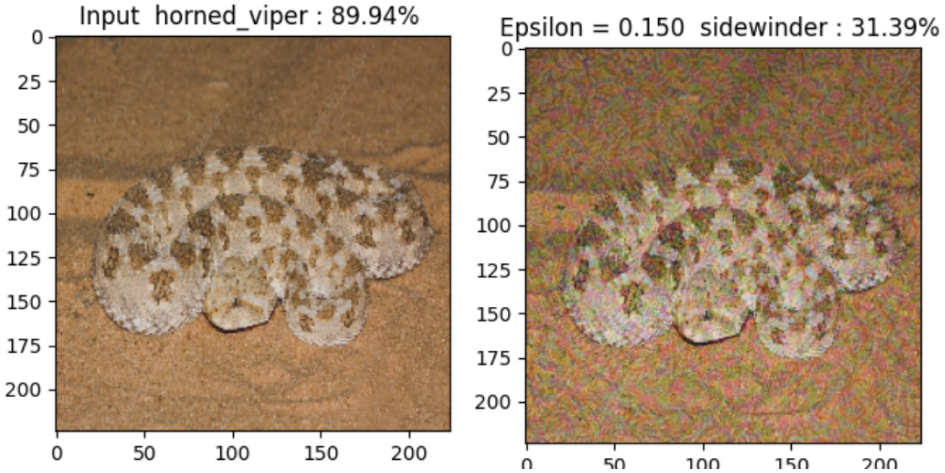

Figure 1: Original (left) vs FGSM Adversarial Image (right)

After that we take the adversarial image and the original image to see the heatmap of Grad-CAM and IG. We used Grad-CAM to visualize the important areas contributing to the classification decision. In the original image, Grad-CAM highlighted the snake's head and body. In the adversarial image, the heatmap became blank and showed no strong focus, which indicates model confusion. And in the Integrated gradient shows the important pixels that are contributing in the decision making.

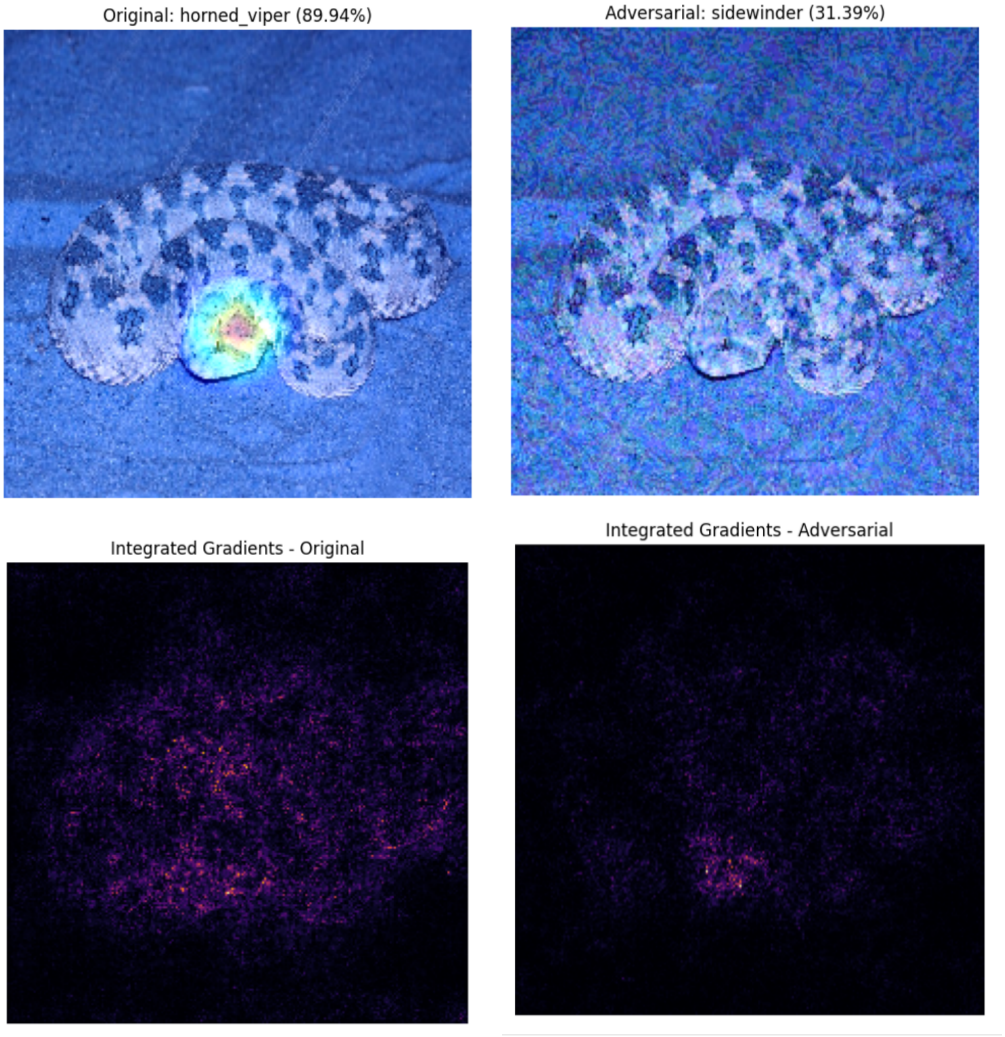

Figure 2: Grad-CAM Heatmaps and IG of Original (left) vs Adversarial (right)

2

In here, we analyzed how the model output with softmax vs. pre-softmax affects two popular model interpretability methods: Grad-CAM and Integrated Gradients (IG). We used a pre-trained MobileNetV2 model and tested both methods on a correctly produced output image of a horned viper.

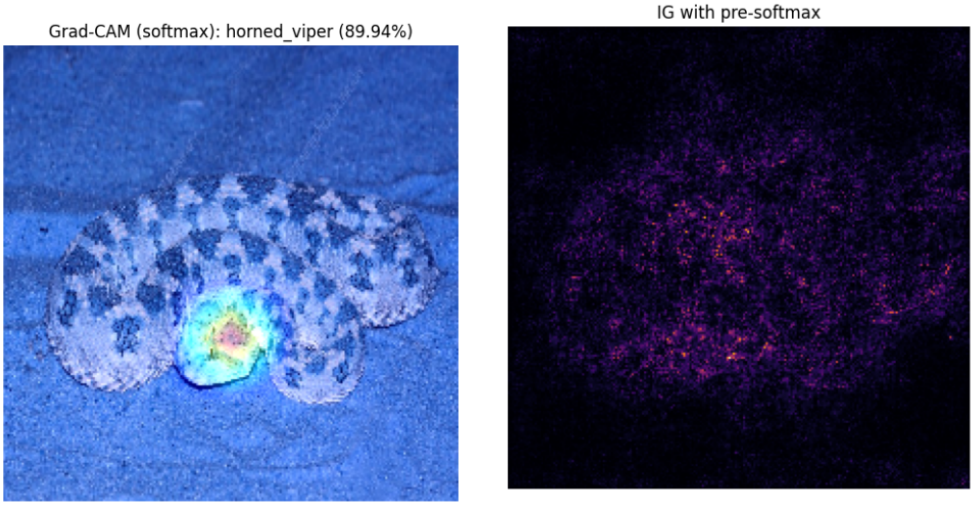

Figure 3: Grad CAM with softmax left and IG pre layers of softmax)

source of colab